A PECULIAR CODON USAGE PATTERN REVEALED AFTER REMOVING THE EFFECT OF DNA METHYLATION

Xuhua Xia

Department of Biology, University of Ottawa
E-mail: xxia@uottawa.ca

Keywords: DNA methylation, deamination, codon usage, genome, genomics

Summary

DNA methylation and deamination increases the C→T mutation rate in CpG dinucleotides, especially in vertebrate genomes. This has profound effect on codon usage in heavily vertebrate genomes, and may obscure the effect of other factors on codon usage bias. We have classified the sense codons into three groups: those decreased by DNA methylation (i.e., CpG-containing codons), those increased by DNA methylation (i.e., TpG- and CpA-containing codons), and those not directly affected by DNA methylation, and studied the codon usage of the last group. RRR and YYY codons are used significantly more frequently than the rest of the codons. This pattern is much stronger in vertebrate genomes than in other genomes and can be used as a content sensor in gene finding.

Introduction

DNA methylation is a ubiquitous biochemical process observed in both prokaryotes and eukaryotes. In vertebrates, DNA methylation is catalyzed by methyltransferases of which a typical representative is the mammalian DNMT1. DNMT1 has five domains of which the NlsD, ZnD and CatD domains bind specifically to unmethylated CpG, methylated CpG and hemimethylated CpG sites, respectively (Fatemi *et al.*, 2001). Methylation of C in the CpG dinucleotide greatly elevates the mutation rate of C to T through spontaneous deamination of the resultant m^5C (Xia, 2003).

DNA methylation should have direct consequences on the evolution of codon usage. The frequency of the CpG-containing codons (e.g., CGA) should be reduced by the joint effect of DNA methylation and spontaneous deamination, and the frequency of TpG-containing and CpA-containing codons, should increase consequently. This has profound effect on codon usage. For example, four codon families (i.e., Ser, Pro, Thr, Ala) have codons with the CpG dinucleotide occupying codon positions 2 and 3. DNA methylation and spontaneous deamination will mediate the change of these NCG codons to NTG and NCA codons, with the former change being nonsynonymous and the latter synonymous. Because nonsynonymous substitution is generally much rarer than the synonymous substitutions in a large number of protein-coding genes in many organisms studied (Xia, 1998a; Xia *et al.*, 1996; Xia and Li, 1998), we should expect NCG to change to NCA more often than to NTG codons. This implies that NCG codons will be underused and NCA codons will be overused in these four codon families.

The effect of DNA methylation may obscure the effect of other factors contributing to codon usage bias. Several factors have already been postulated to affect codon usage. For example, tRNA molecules carrying the same amino acids often differ much in concentration and an increase of the synonymous codons matching the anticodon of the most abundant tRNA would increase the translation rate (Bulmer, 1991; Ikemura, 1992; Xia, 1998b). Although these two hypotheses can account for much variation in codon usage patterns, there is still much unexplained variation, much of which might be caused by the confounding effect mediated by DNA methylation and spontaneous deamination.

One method to remove the effect of DNA methylation on codon usage is simply to classify the sense codons into three groups, one including codons whose frequencies would be reduced by DNA methylation (i.e., all CpG-containing codons), one including the codons whose frequencies would be increased by DNA methylation (i.e., all TpG- and CpA-containing codons) and the third including all other codons whose frequencies are not directly affected by DNA methylation. The codon usage pattern in this third group should not be confounded by the effect of DNA methylation on codon usage and may consequently reveal codon usage patterns not observed previously.

Materials and Methods

We have used the following vertebrate representatives: *Homo sapiens, Xenopus laevis*, and *Danio rerio*, and the following non-vertebrate representatives *Drosophila melanogaster, Caenorhabditis elegans,* and *Saccharomyces cerevisiae.* The coding sequences were extracted from UniGene files (xx.seq.uniq, where "xx" stands for the species abbreviation) available at ftp.ncbi.nih.gov/repository/UniGene for the following species: *Homo sapiens, Xenopus laevis*, and *Danio rerio.* The coding sequences for *D. melanogaster* were retrieved from scaffold data (AE00xxxx.ffn) available at the FTP site (ftp.ncbi.nih.gov/genbank/genomes). Only results from the first scaffold (AE002566) with 1191 CDS sequences was presented because the codon usage patterns from the other 18 scaffolds are all very similar. The *C. elegans* data were retrieved from the same FTP site. Only results from the first *C. elegans* chromosome was presented because the codon usage patterns from the other chromosomes are similar. The coding sequences for *Saccharomyces cerevisiae* were retrieved from Saccharomyces Genome Database at http://genome-www.stanford.edu/Saccharomyces. Only complete CDSs starting with a methionine codon and ending with a termination codon were used.

Codon frequencies depend on nucleotide frequencies, e.g., GC-rich DNA should have more GC-rich codons. Furthermore, the nucleotide frequencies typically are different at different codon position so that even a CDS sequence with equal number of A, C, G, and T may have codon frequencies differing much from each other. For example, if G occurs only at the first codon position and C only at the second codon position, then we would have many GCN codons but few CGN codons. Therefore, we have compiled the observed codon frequencies and computed the expected frequencies from nucleotide frequencies for each codon positions. Designating observed codon frequencies as O and expected as E, we computed the (O-E)/E which is termed SR (for standard residue) hereafter. We expect the codons in the FROM group to have small SR values and those in the TO group to have large SR values.

The counting of codon frequencies was done in two ways, one including the initiating codon and the other excluding the initiating codon. The conclusions presented in this paper are valid for both counting procedures, and only the result from the counting procedure excluding the initiating codon is presented in tables. The expected codon frequencies were calculated as follows. Let P_{ij} be the site-specific nucleotide frequencies of the sense codons, where i = 1, 2, 3 corresponding to codon positions 1, 2, and 3, and j = 1, 2, 3, 4 corresponding to A, C, G and T, respectively. The expected frequency of codon AGA is then

$$P_{AGA} = \frac{P_{1,1}P_{2,3}P_{3,1}}{SumP} \tag{1}$$

where SumP is the sum of the $P_{1i}P_{2j}P_{3k}$ terms for all sense codons. For the universal genetic code, SumP is the sum of 61 $P_{1i}P_{2j}P_{3k}$ terms. The calculation of P_{ij} did not include stop codons because the inclusion would underestimate the frequency of C. The computation of site-specific nucleotide frequencies, the expected codon and di-codon frequencies, and all subsequent statistical analyses were done with DAMBE (Xia, 2001; Xia and Xie, 2001).

Results and discussion

The nucleotide frequencies for each of the three codon positions were counted for the 15109 human CDS sequences. These nucleotide frequencies were then used to compute the expected frequencies of the 61 sense codons for the human CDS sequences. The observed frequencies of the 61 sense codons (Table 1) are largely consistent with the effect of methylation on codon usage. The eight CpG-containing codons all have observed frequencies much lower than their respective expected frequencies. The SR value (Table 1) measures the deviation of the observed value from the expected value and is independent of sample size. The mean SR value for the eight CpG-containing codons is -0.4823. In contrast, most of the 15 UpG-containing and CpA-containing codons have their observed frequencies higher than their respective expected frequencies, with the mean SR value equal to 0.2022.

Table 1. The observed (O) and the expected (E) frequencies for the 61 sense codons classified into three groups, with IDs of "-1", "1", and "0" for codons decreased, increased and not affected by DNA methylation, respectively. SR = (O-E)/E. Based on 15109 human CDS sequences. The "1" group does not include the UG-containing UGA codon because it is not a sense codon.

Codon	*O*	*E*	*SR*	*ID*	*Codon*	*O*	*E*	*SR*	*ID*
ACG	44084	133005.8	-0.669	-1	CCC	144034	128711.1	0.119	0
CCG	51567	122948.6	-0.581	-1	AGA	87107	73850.7	0.180	0
CGA	45717	68266.5	-0.330	-1	GGC	163689	134313.1	0.219	0
CGC	76261	104560.8	-0.271	-1	GCC	202803	165335.3	0.227	0
CGG	83435	99879.5	-0.165	-1	AGC	140972	113114.0	0.246	0
CGU	33777	79077.3	-0.573	-1	AAG	233830	177648.4	0.316	0
GCG	54254	157933.0	-0.657	-1	CCU	128978	97341.7	0.325	0
UCG	32683	84626.2	-0.614	-1	GAG	291059	210942.2	0.380	0
GUA	52660	120106.6	-0.562	0	GGA	122104	87691.4	0.392	0
AUA	54170	101149.7	-0.465	0	UCC	125479	88592.6	0.416	0
CUA	51936	93501.3	-0.445	0	UUC	144076	98573.5	0.462	0
GUC	103360	183962.1	-0.438	0	AAA	183074	121420.6	0.508	0
GUU	80933	139126.9	-0.418	0	GAA	220405	144176.6	0.529	0
AAC	139177	185974.7	-0.252	0	UCU	109113	67000.8	0.629	0
AGG	83842	108049.7	-0.224	0	UUU	125983	74549.2	0.690	0
GGU	79894	101578.4	-0.214	0	ACA	109339	90908.0	0.203	1
GAC	185126	220829.0	-0.162	0	AUG	143532	147990.4	-0.030	1

UUA	56384	64357.4	-0.124	0	CAA	91270	112239.4	-0.187	1
CUU	95737	108308.3	-0.116	0	CAC	108309	171912.2	-0.370	1
AAU	125990	140649.0	-0.104	0	CAG	250071	164215.5	0.523	1
ACU	96179	105304.3	-0.087	0	CAU	79270	130013.8	-0.390	1
GGG	118196	128299.8	-0.079	0	CCA	125584	84033.9	0.494	1
UAC	109345	118328.1	-0.076	0	CUG	283383	136800.1	1.072	1
CUC	137174	143211.9	-0.042	0	GCA	117312	107945.4	0.087	1
AUC	148829	154926.7	-0.039	0	GUG	203095	175725.9	0.156	1
ACC	135017	139239.8	-0.030	0	UCA	88625	57841.0	0.532	1
GAU	164887	167008.6	-0.013	0	UGC	89379	71969.8	0.242	1
UAU	88939	89489.2	-0.006	0	UGG	90399	68747.6	0.315	1
AUU	117197	117168.0	0.000	0	UGU	76049	54429.3	0.397	1
AGU	90301	85545.9	0.056	0	UUG	93206	94160.3	-0.010	1
GCU	135067	125039.8	0.080	0					

The mean SR value for those codons not affected by DNA methylation is intermediate, being 0.0494. Significant differences exist among the three means (ANOVA and multiple comparisons with $P < 0.0001$), suggesting a significant effect of DNA methylation on the usage of sense codons. Note that this effect of methylation on codon usage would be much obscured if we looked at only the observed frequencies without comparing them to the expected codon frequencies (Table 1) based on the nucleotide frequencies at the three codon positions.

The 38 codons not directly affected by DNA methylation were sorted by the SR value, which reveals a novel pattern of codon usage (Table 1). The RRR and YYY codons are used more frequently than the rest of the codons, and this pattern would have been obscured if we did not eliminate the methylation effect because the UpG-containing codons are non-RRR and non-YYY but generally have large SR values.

Designating the 16 RRR and YYY codons as Group 1 and the other 22 codons as Group 2, we found highly significant difference in the mean SR value between the two groups (Table 2). The difference is true not only for the vertebrate species in Table 2, but also for a number of other vertebrate species we tested, including *Mus musculus, Rattus norvegicus, Bos Taurus, Gallus gallus, and Alligator mississippiensis*. The difference is smaller for non-vertebrate species, and is not significant for *D. melanogaster*. However, the direction of the difference, even for *D. melanogaster*, is consistent.

Table 2. T-tests of whether RRR and YYY codons are used more frequently. Mean1 – Mean SR for all RRR and YYY sense codons, Mean2 – Mean SR for codons other than RRR and YYY. The degree of freedom is 36 for all tests.

Species	N_{CDS}	Mean1 (SE)	Mean2 (SE)	T	P
H. sapiens	15109	0.2802 (0.0695)	-0.1184 (0.0495)	4.8068	<0.0001
X. laevis	1555	0.2698 (0.0642)	-0.1306 (0.0383)	5.6645	<0.0001
D. rerio	900	0.2285 (0.0952)	-0.1380 (0.0491)	3.6896	0.0007
C. elegans	2474	0.1595 (0.1313)	-0.1929 (0.0486)	2.8143	0.0079
D. melanogaster	1191	-0.0097 (0.1204)	-0.0336 (0.0660)	0.186	0.8535
S. cerevisiae	6357	0.1308 (0.0681)	-0.0565 (0.0517)	2.2325	0.0319

Our results suggest that that DNA methylation can contribute substantially to codon usage bias. Taking DNA methylation in account should greatly increase the explanatory power of the two existing hypotheses, i.e., the transcriptional (Xia, 1996) and the translational (Bulmer, 1991; Ikemura, 1992; Xia, 1998b) hypotheses on codon usage bias.

Acknowledgement

This study is supported by grants from University of Ottawa and NSERC of Canada.

References

Bulmer, M., 1991, The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129: 897-907.

Fatemi, M., Hermann, A., Pradhan, S., and Jeltsch, A., 2001, The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA. *J. Mol. Biol.* 309: 1189-1199.

Ikemura, T., 1992, Correlation between codon usage and tRNA content in microorganisms. In *Transfer RNA in protein synthesis.* Hatfield, D.L., Lee, B. and Pirtle, J. (eds). CRC Press, Boca Raton, Fla., pp. 87-111.

Xia, X., 1996, Maximizing transcription efficiency causes codon usage bias. *Genetics* 144: 1309-1320.

Xia, X., Hafner, M.S., and Sudman, P.D., 1996, On transition bias in mitochondrial genes of pocket gophers. *J. Mol. Evol.* 43: 32-40.

Xia, X., 1998a, The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. *Mol. Biol. Evol.* 15: 336-344.

Xia, X., 1998b, How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? *Genetics* 149: 37-44.

Xia, X., and Li, W.H., 1998, What amino acid properties affect protein evolution? *J Mol Evol* 47: 557-564.

Xia, X., 2001, *Data analysis in molecular biology and evolution.* Kluwer Academic Publishers, Boston.

Xia, X., and Xie, Z., 2001, DAMBE: Software package for data analysis in molecular biology and evolution. *J. Hered.* 92: 371-373.

Xia, X.H., 2003, DNA methylation and mycoplasma genomes. *J. Mol. Evol.* 57: S21-S28.